Abstract

Open Self-Medication\(^1\) is a Web application that better informs people when treating undiagnosed medical ailments with unprescribed, over the counter drugs, \textit{i.e.}, self-medicating. The application achieves this goal by providing a set of functionalities that ensure safety and efficiency of this practice. The system’s most critical operations are processed using a self-medication knowledge base, expressed in OWL, which has been inductively built on medical information obtained from a similar French project. A main characteristic of this application is that almost all the data processed by the system and presented to the end-user comes from a subset of the LOD data sets, namely \textit{DrugBank}, \textit{DailyMed}, \textit{Sider} and \textit{DBPedia}. This paper motivates the design of such an application, provides the design choices, describes some implementation details and presents lessons learned and future work.

\textit{Keywords:} Linked open data, knowledge base, OWL, self-medication, web application

1. Introduction

In OECD (Organization for Economic Co-operation and Development) countries, due to the emergence and wide distribution of new drugs as well as the aging of populations, the consumption of pharmaceuticals keeps increasing every year. In 2009, the associated bill has been estimated to USD 700 billion, \textit{i.e.,} 19\% of the total health spending [12]. On the one hand, this practice comes at an increasing financial cost for the patient and the health
system and may also imply some health risks, \( e.g. \), contraindications between molecules taken during a given period of time, overdosing a molecule present in different drugs. On the other hand, this rising consumption of drug products is not necessarily correlated with the general public’s access to medical information that helps them make more informed decisions. A common practice nowadays is that the general public totally relies on the skills of health care professionals when it comes to consuming drugs.

Meanwhile, in the same OECD document, it is reported that self-medication (\( i.e. \), the act of treating undiagnosed medical ailments with unprescribed drugs) with over-the-counter (OTC) pharmaceutical products typically accounts for approximately 15% of drug spending. This has been exacerbated since 2008, coinciding with the beginning of the financial crisis, with people trying to reduce their health related spending, \( e.g. \), avoiding visits to general practitioners whenever possible. This has contributed to the fact that self-medication is one of the most dynamic drug markets for pharmaceutical companies. Most experts of the domain consider that this practice will accentuate in the coming years. Industrialized countries like France are pushing in this direction by transitioning molecule sets from originally prescribed to OTC, \( i.e. \), opening them to self-medication. This political approach can have a disastrous impact on the health of the general public if not guided properly [13].

Due to the increasing practice of self-medication, many government agencies are calling for the development of educational tools targeting the general public on health and medical issues. At the same time, other government agencies, usually in the same countries, are requesting to implement computerized applications based on open data. The Open Self-Medication project aims to satisfy both calls by developing a self-medication Web application fueled on Linked Open Data (henceforth LOD) data sets dedicated to drugs and medicine.

The application presented in this paper achieves this goal by providing a set of functionalities that ensure safety and efficiency of the self-medication practice. With safety, we mean that the system guides the end-user from a set of common mild medical signs to adapted molecules and drug products, but also highlights the risks, \( e.g. \), drug interactions, adverse events, of self-prescribing a drug in a given situation. The efficiency argument corresponds to providing a rating, based on a tolerance/efficiency ratio designed by a team of health care professionals, to some identified self-medication molecules. This project leans on the experience of implementing a similar
system for the French market. Together with Pr. Jean-Paul Giroud (PhD, MD, former pharmacology WHO expert), we have contributed to the practice of a safe and efficient self-medication by releasing several books, recently [7], and Web applications, one of which has been accessible for the last 6 years to over 6 million clients of three major insurance companies in France. This application, named Les conseils de l’automédication, is owned by the Santéclair service company and is only granting access to its clients. Recently, an iOS application, namely Top des médics, has been released and is accessible to a larger audience, i.e., those purchasing the application.

Intuitively, the French application, influenced by [4], aims to provide adapted to the general public, objective (i.e., not influenced by pharmaceutical companies), information on drug products sold in France. In just a few interactions, one can select a symptom and obtain information on associated therapeutic classes and drugs. A main feature of this application is the rating of drugs, i.e., a typical French score between 0 and 20, based on an efficiency/tolerance ratio. This ratio depends on the drug’s composition, i.e., its set of molecules, as well as its indication. The cornerstone of the system is a drug database which is the result of years of medical and pharmacological research, e.g., translating drug characteristics into chunks of text that are easily understandable by the general public. The back end of the Web application is based on the use of a knowledge base which has been designed and enriched using inductive reasoning over the drug database [6]. The ontology that underpins this knowledge base also serves to ensure the quality of the drug data by performing molecule-oriented inferences [5]. That is reasoning over the properties, e.g., contraindications, of molecules contained in a drug rather than on the properties of a drug. This approach enables to identify drugs in the scope of self-medication by reasoning over the molecule it is composed of.

Unlike the system we developed previously for the French market, the Open Self-Medication application more heavily uses data originating from the open data initiative. Hence, we believe it will be of practical value to a more general audience of Web users since it enables to retrieve information from different health systems. To achieve this goal, we had to define correspondences between molecule identifiers found in our select LOD sets and the ones already present in our Knowledge Base. Our main motivation is to leverage open data, as represented in Semantic Web technologies, and to apply our experience with the French system toward a self-medication application that will better serve an international audience, i.e., support for
drug products sold in different countries and providing their description in different languages. Moreover, the goal of using open data amounts to accessing large volumes of data which are hard to obtain in the proprietary pharmacological domain.

2. Related work

The design of a self-medication application is relatively novel and does not currently correspond to any existing work. Nevertheless, related work exist on the topic of exploiting LOD in domain such as life sciences and medicine. As a motivation for our approach, [9] considers that using open data is one of the most efficient ways to collect data and knowledge on the medical domain at large scale. Moreover, as explained in [11], semantic technologies will prove to be amongst the most reliable methods to search into these data sets. In [8], the authors motivate the technical aspects selected during the creation the Open Pharmacological Space (OPS), a drug discovery platform part of the Open PHACTS project. Although the principal goal of this application differs from ours, i.e., helping in drug discovery versus providing accurate drug information on known products, the approaches share some common aspects such as exploiting data sets of the LOD cloud. We also share opinions on the data quality issue of some of these data sets. [14] presents the Linking Open Drug Data (LODD) W3C’s task force and its project of linking drug related data, some of which have been used in this project. It also proposes best practices for exposing data that could fit efficiently into LOD. Finally, [3] describes an experience in developing a food ontology in the context of the FP6 PIPS (Personalized Information Platform for Health and Life Services) project. The ontology development process did not rely on an inductive approach like Open Self Medication and did not use external knowledge repositories as most recent approaches do.

3. LOD data sets

Some design choices concerning this application were mainly motivated by the results of investigating the LOD data sets relevant to the domain of self-medication. The data sets we formally considered were DrugBank[10],
We finally retained only the first four since we considered that the others were not adapted to self-medication, e.g., diseasesome and LinkedCT, or were less precise than some selected data sets, e.g., FreeBase vs DBPedia.

An in-depth study of our selected data sets rapidly emphasized that the central concept of this application’s first version would not be the same as in our French system. The central concept in the French application corresponds to a Drug product for the following reasons: (i) a unique identifier, namely the CIP (Club Inter Pharmaceutique) code, is provided to each drug product being sold in France, (ii) information sources are available, although distributed over several providers, e.g., ANSM (Agence Nationale de Sécurité du Médicament et des produits de santé), on these products and (iii) market evolution on these products are also available, i.e., modification, emergence and withdrawal of products. Thus it is possible to maintain an almost exhaustive, up-to-date and accurate drug database for these drugs. Note that (i) and (ii) are characteristics of different national health systems and are not consequences of LOD per se. These three properties also enable us to rate drugs and molecules in a consistent manner. The ability of rating molecules is quite useful in cases of compound drugs containing several molecules with different dosages. For instance, our health experts can use the rating of some molecules to help in defining the rating of a drug composed of several molecules. In the case of drugs found on LOD, we have not found an accurate and consistent list of identified drugs available on the market at a given time. For instance, at the time of developing this application, data sets such as RXNorm are not present on the LOD cloud. For this reason, we have currently decided to only rate molecules and to consider non-compound drugs. Moreover, drug products containing a single molecule generally have better properties, i.e., less drug interactions, contraindications, side-effects, and hence have higher rates and provide a better service to the patient. Nevertheless, we are already working on compound drugs since informing the general public about inefficient or/and dangerous drug products should also be an important feature of our system.

---

2http://dailymed.nlm.nih.gov/dailymed
3http://sideeffects.embl.de/
4http://linkedct.org/
5diseasome.eu/
4. Self-Medication Knowledge Base

The issue of selecting the molecules fitting in the field of self-medication is non-trivial. This is partly due to the absence of an explicit ‘self-medication’ indication in most drug data and knowledge bases. In fact, we consider that being an OTC drug is not a sufficient condition to belong to our self-medication category. For instance, in France, some drugs belong to the family of OTC products but we do not recommend them for self-medication due to some of their properties, e.g., contraindications or unadapted therapeutic classes.

The identification of molecules in the scope of self-medication is performed with the assistance of a knowledge repository. This knowledge base, developed in the context of our French application, has been designed semi-automatically using induction over our French drug database. This knowledge repository regroups valuable information on molecules which are used by our team of health experts to qualify if they belong or not to the self-medication domain. We now present the main steps involved in developing this knowledge base and invite interested readers to obtain more details in [6]. The method is able to generate an OWL ontology from any drug related classification present in our drug database. In a nutshell, our self-medication database stores, for all drugs sold in the French market, all the information available on the Summary of Product Characteristics (SPC). In addition, it also contains our efficiency/tolerance ratio, and a comment from a team of health care professionals. This database takes the form of a star-shaped schema where the central table stores information on general drug information, e.g., name, price, as well as CIP drug identifiers which serve as primary keys. This table is also related, using one-to-many or many-to-many associations of the Entity-Relationship logical model, to all kinds of medical information, e.g., side-effects, therapeutic classes, contraindications, warnings. Drugs are also related to some of the most frequently used, at least in Europe, drug classifications, e.g., ATC (Anatomical Therapeutic Chemical Classification System)\(^6\) and EphMRA (European Pharmaceutical Market Research Association)\(^7\). Let’s concentrate on ATC, it is organized in a five-level hierarchy ranging from anatomical groups to molecules that corresponds to a generalization/specialization classification. It uses an encoding system that

\(^6\)http://www.whocc.no/atcddd/
\(^7\)http://www.ephmra.org/
can easily be transformed into an explicit OWL concept subsumption hierarchy, e.g., a R5D concept, with an Antitussives label, is created and has the R5 concept, with label Respiratory System as a super concept. This approach, when performed over ATC, yields a concept hierarchy of depth 5 containing over 5,000 concepts. Starting from this concept hierarchy, our method associates properties, corresponding to some medical information, to each concepts corresponding a molecule. This is performed using the star-shaped schema to cluster relevant groups of products, generated using one of the previously cited classifications. Intuitively, we navigate in the hierarchy of concepts and create groups of products for each level, using the classification to drug relation in our database. Then, for each group we study some specific properties which correspond to fields in SPCs (Summary of Product Characteristics), e.g. contraindications, and for each possible value in these domains we calculate the ratio of this value’s number of occurrences on the total number of elements of the group. We attribute the property, i.e., via a previously defined OWL property, to a concept when that ratio is above a predefined threshold. This step is semi-automatic, meaning that a health care professional supervises the results inferred and has the ability to overrule them. This is performed with the help of a graphical user interface and hence is processed efficiently by our health experts by usually requiring only a small numbers of mouse clicks. Finally, we refine the ontology by moving property assertions along a subsumption branch. This is performed automatically to remove redundancy among siblings by moving properties up in the hierarchy when they are shared among all children of a given concept. Thus this ontology refinement approach is deterministic and based on its topology.

In order to select self-medication relevant molecules from our LOD data sets, our first approach was based on identifying molecules with the ATC or EphMRA classifications. We rapidly found out that these identifiers are sparsely used in DrugBank and DBPedia, the central data sets of our application. So we complemented our ontology with a new identifier solution based on DrugBank keys. This processing step was performed semi-automatically using different sources such as DBPedia and other classifications, e.g., EphMRA. At this stage, we also found out that some molecules could not be matched because they could not be found due to their non-commercialization in North America, e.g., absence of the Ambroxol molecule. We found out that most missing drugs were badly rated and considered useless in our French system.
5. System description

Figure 1 provides an overview of the main components of the system, i.e., LOD data sets, database instance and knowledge base. It highlights that the communication between the different data repositories is performed using a set of APIs and interfaces. Next we consider the graphical user interface which provides the medical information to the end-user.

In Figure 2, we present a screenshot of the application that displays information on the Dextromethorphan molecule. The navigation bar (upper part of the screen) proposes several functionalities, i.e., home, symptoms, molecules, drugs, contact us and find a pharmacy. The 'symptom' feature enables the end-user to access molecules and drugs from mild clinical signs. Given our approach within the design of the French system, we consider that this is the most intuitive way for the patient to find some valuable information. This feature enables to design a classical self-medication scenario: the end-user selects an entry among the most frequent self-medication related symptoms, e.g. coughing, a list of adapted molecules is then displayed with short indication and its rating (from A to E). More information, e.g., contraindications, side-effects, food interactions and drugs, can be obtained from selecting a molecule. Finally, among the list of drugs, some of them, i.e., those that provide useful and relevant information in a self-medical context, can be selected to get more information, i.e., adverse events, precautions, warnings. The list of molecules displayed in the Interactions section are linked, thanks to URIs and the linking of data sets, to drug products stored in DBPedia and Drugbank. Hence, clicking on a molecule name opens up a box that contains all drugs containing it. This is quite useful for most end-users who generally know some product names but do not know the molecules they are composed of. Because we were not satisfied with the amount of drug information one can retrieve from LOD data sets, the molecule information page provides a link to DrugBank pages which contain more information, e.g., drug prices.

The navigation bar also proposes a more direct access, i.e., not implying the selection of symptom, to molecule and drug information. This approach enables to inform the general public on non self-medication molecules and drugs by proposing to access information on all drugs and molecules obtained on our selected LOD data sets. Up to now, we do not provide ratings for these molecules. This is mainly motivated by the fact that they are generally out of the scope of self-medication.
Figure 1: Architecture overview

**DrugBank**
detailed data on 6729 drugs

**DailyMed**
product labeling for marketed drugs in the US

**Sider**
data on indications and side-effects

**DBPedia**
data from Wikipedia on drugs

**LOD sources**

**APIs and Interfaces**

**Heartburn**
Heartburn, also known as pyrosis, cardialgia, or acid indigestion is a burning sensation in the chest area behind the breastbone or in the epigastrium, the upper central abdomen. The pain often rises in the chest and may radiate to the neck, throat, or angle of the jaw.

**MOLECULES FOR HEARTBURN**

- **Famotidine**: rated A
  - For the treatment of peptic ulcer disease (PUD) and gastroesophageal reflux disease (GERD).

- **Cimetidine**: rated C
  - For the treatment and the management of acid-reflux disorders (GERD), peptic ulcer disease, heartburn, and acid indigestion.

- **Pantoprazole**: rated B
  - Short-term (up to 16 weeks) treatment of erosive esophagitis.

- **Nizatidine**: rated C
  - For the treatment of acid-reflux disorders (GERD), peptic ulcer disease, active benign gastric ulcer, and active duodenal ulcer.

**User-facing Information**

MySQL KB

drug ratings
drug hierarchies
pharmacies
Finally, to provide useful services in the context of self-medication, we propose the end-user to locate the pharmacies surrounding her current or a defined location (Figure 3). Given a position, this tool suggests the ten closest pharmacies and provides driving/walking directives. We have not found on LOD a repository of addresses for health actors, e.g., hospitals, doctors, pharmacies. Hence, we harvested the Web for pharmacy coordinates in France and the US. Currently, we have gathered more than 18,000 of them. At the moment, these information are stored in a relational database, i.e., MySQL, since it does not follow a graph representation and enables fast data retrieval due to indexes created on the geolocation coordinates.

6. Implementation details

The Web application uses HTML5 and CSS3 for the user interface and programs are written using PHP, Javascript and its JQuery library. The ge-
Concerning the LOD data sets, we are not using SPARQL endpoints due to their relative unreliability. Instead, we downloaded some dumps and put them on our own OpenRDF Sesame triple store. We are using the phpSesame PHP Client library which enables to access OpenRDF’s Sesame Framework via HTTP requests. All non RDF data, i.e., list of mild clinical signs, molecules and their ratings as well as pharmacy coordinates, are stored in a Mysql database instance to support a faster and easier integration with our French system. A stored procedure enables to compute the distance...
between a given two pairs of latitude/longitude coordinates, \textit{i.e.}, the user-end
and pharmacy locations. Retrieving data from these data stores is performed
using REST services which are using SPARQL and SQL queries. We are
aiming to propose some public APIs for these services in order to enable
interested developers to access them directly.

7. Lessons learned

The main challenge addressed by this work is to evaluate the feasibil-
ity of designing a self-medication application relying on the LOD. Open
Self-Medication is a proof of concept that such an application can be im-
plemented and, considering positive feedbacks from end-users, can provide a
useful service to the general public. The main contributions consist in the
identification of a subset of the LOD to self-medication and the selection rel-
evant molecules. The latter leverages on research previously conducted over
several years on a similar French system, \textit{i.e.}, the design of a self-medication
knowledge base and its use in enhancing data quality, \cite{5} form more details.

Considering the application and domain point of views, we consider that
the use of the LOD provides many benefits. First, it enables to address
a large audience with the description, in different languages, of drugs sold
in different countries. Characteristics of LOD also support the discovery
of new data sets and valuable information by following dereferenceable URIs.
For instance, it enables to integrate data in an efficient and fluid way that
could not be supported with non RDF formats. In terms of other usage of
semantic technologies, reasoning over our knowledge repository is quite useful
but is mainly supporting back end operations, \textit{e.g.} data quality enhancement.
Finally, we appreciate the ability to query in a federation kind of way over
several LOD data sets, \textit{i.e.}, executing a SPARQL query over several RDF
graphs.

The development of this application also emphasizes the heterogeneity
and data quality issues of the data contained in some LOD data sets. In terms
of heterogeneity, some molecules are lacking some important properties, \textit{e.g.},
\textit{ATC} or \textit{DrugBank} molecule identifiers, and are not properly linked to other
data sets. Although some drug products are well documented, others are
totally absent. At the same time, some important properties are missing
for some popular drugs. The solution we have implemented to discover and
filter unsufficiently described molecules is supported by the manual design of
a set of SPARQL queries that were executed over the selected LOD data sets.
These queries were mainly checking the presence of fundamental information, which are selected by health experts from our ontology, the molecule must provide. We consider that 25% of the self-medication molecules present in our knowledge base were rejected following that approach.

It was not a surprise to discover that the content on diseases, molecules and drugs on data sets such as LOD are mainly targeting the community of health care professionals. Studying them, we found out that some information are still comprehensible for the general public but several processing steps are needed to adapt the content to certain categories of end-users. Finally, to the best of our knowledge, there does not exist a free, open source repository of addresses for health care actors and care facilities. Such data sets are highly desirable for developers aiming to implement applications providing directions to hospitals, clinics, pharmacies, and so on.

8. Conclusion and Future works

The Open Self-Medication application is built using Semantic Web technologies, i.e. RDF (through LOD data sets), OWL and SPARQL. Its goal is to provide useful information to the general public on mild clinical signs and self-medication molecules. We consider that this application already provides useful information to the general public on mild clinical signs and self-medication molecules. The use of technologies available in the Web of Data has proved to be quite useful and presents a lot of potential. Nevertheless, we were disappointed with the correctness and completeness of drug information available on the LOD cloud, which limits the breadth of our application. In order to deal with these issues, we have decided to show only those drugs that satisfy a certain quality, i.e., given in terms of correspondences with our own ontology.

Right now, updating the data repository requires significant effort and we aim to automate this process. We also hope to obtain more user feedback to improve the system and to develop new features. In terms of future work, one project could be to adapt a similar tool for health care professionals as primary users, since most of the LOD content seems oriented toward this audience already. This application would of course tackle all kinds of drugs and would not be limited to self-medication and OTC drugs. We already know from our previous French system that many general practitioners are using either text books or applications that we have produced on self-medication. Thus, the goal remains to engage patients more directly as consumers. The
most obvious strategy would be to extend our existing French system into a multilingual version, but language is not the only issue to consider — drugs are approved by different national organizations, bought and used differently across countries and a useful system as to address all these aspects.

References


